UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CS 598 Machine Learning for Signal Processing

Probability, Statistics & Parameter Estimation

28 August 2015

Logistics

- Did everyone get the class email?
  - If not, send me your NetID so that I can add you to the mailing list

- Is there a waiting list to register for the class?
  - Sorry no, just keep trying to register

- Class recordings are available for registered students at:
  - https://recordings.engineering.illinois.edu:8443/ess/portal/section/242d0f51-7fa8-49d2-aa4c-b2b78701dc10
    - Remember attendance counts!

Today’s refresher

- Probability
- Statistics
- Parameter Estimation

Probability

- Probity
  - Measure of legal authority/nobility
    - Passed muster in the middle ages
- Probability
  - Measure of belief/likelihood
    - Passes muster today

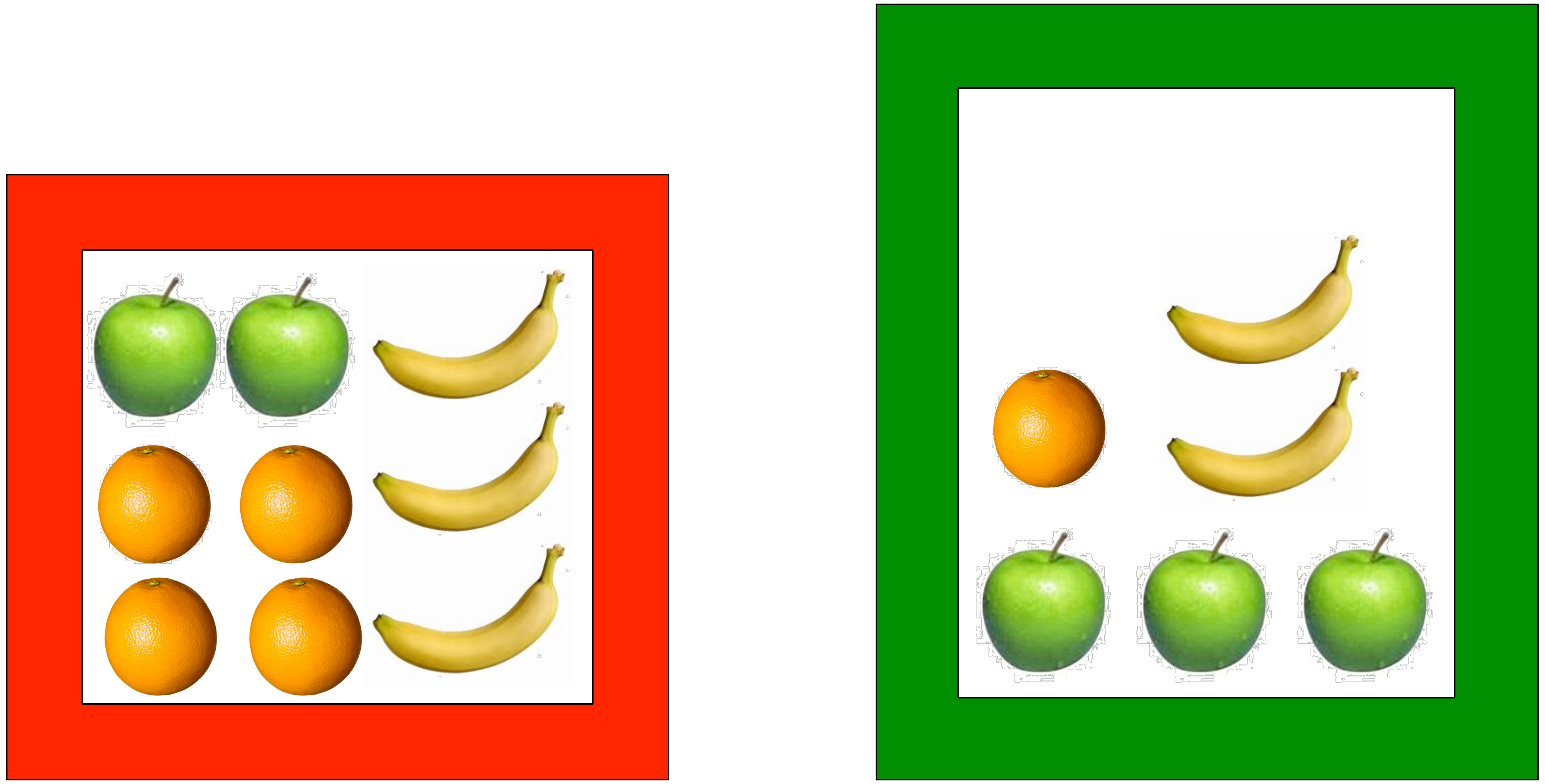
Goals of probability

- Characterize stochastic processes
  - How do dice roll?
  - What am I more likely to say next?

- Indicate belief given evidence
  - The suspect was nearby and there are feathers on his clothes. Was he the chicken thief?

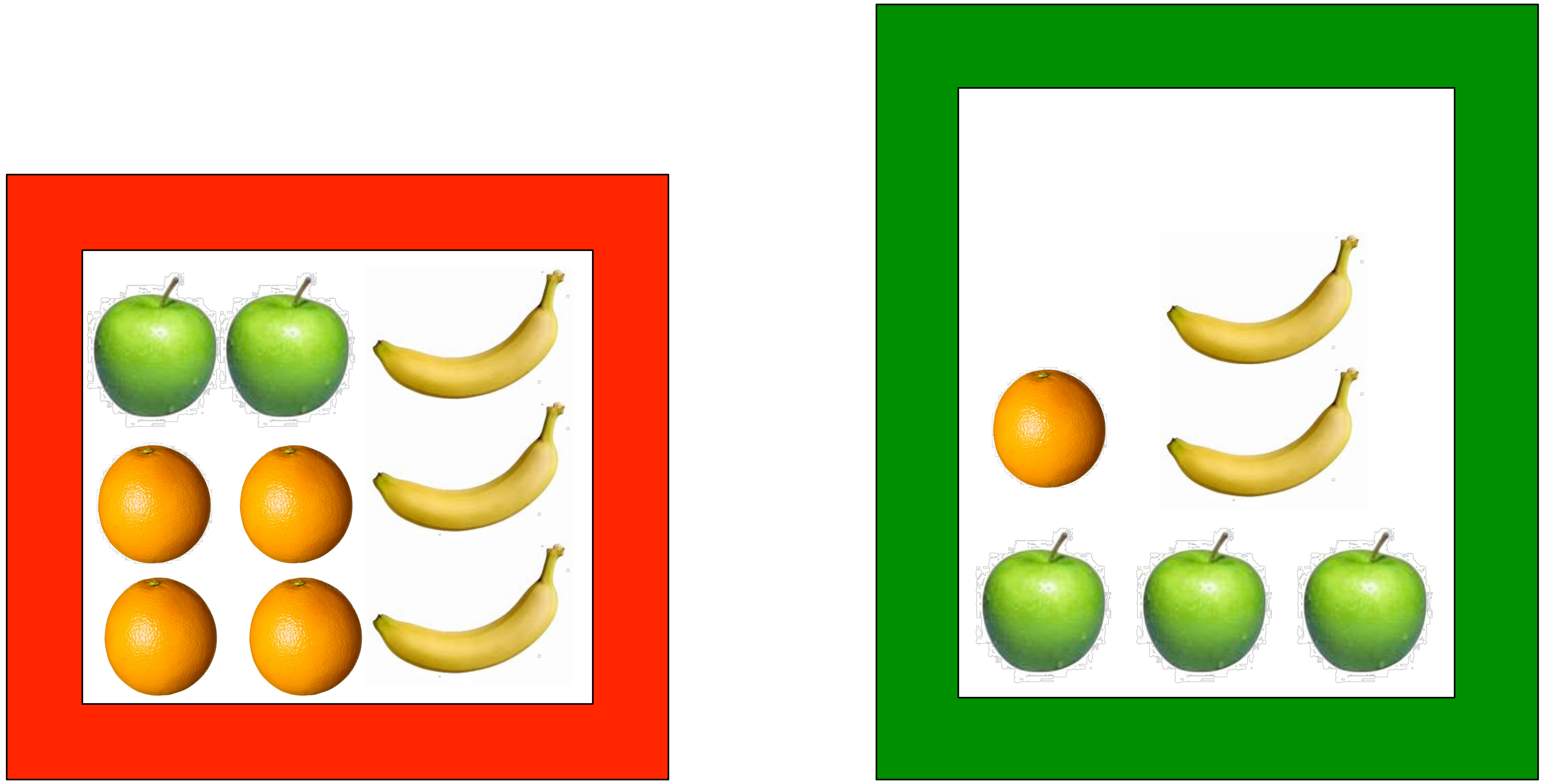
An example

- We start picking oranges, apples and bananas, from the two boxes below
  - Pick 40% from red box, 60% from green box



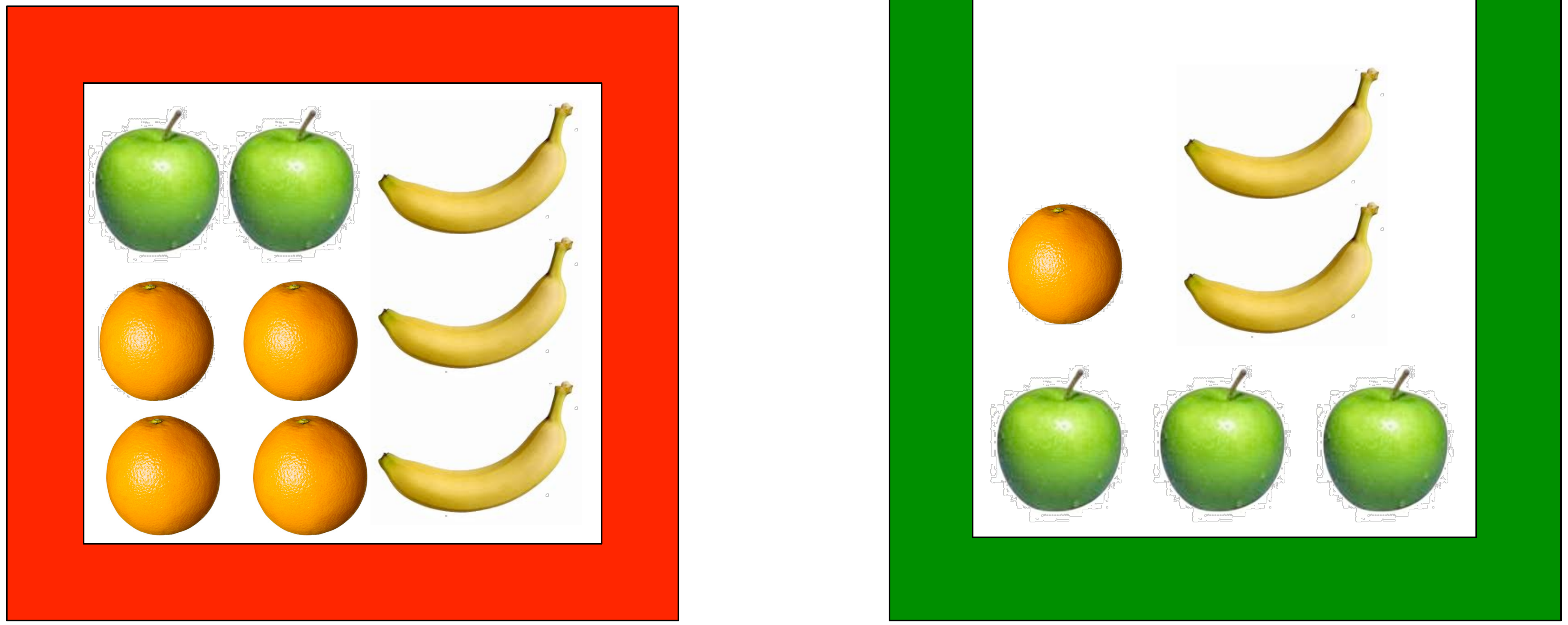
The random variables

- The box: B = {r, g}
- The fruit: F = {a, o, b}
  - What are their probabilities?



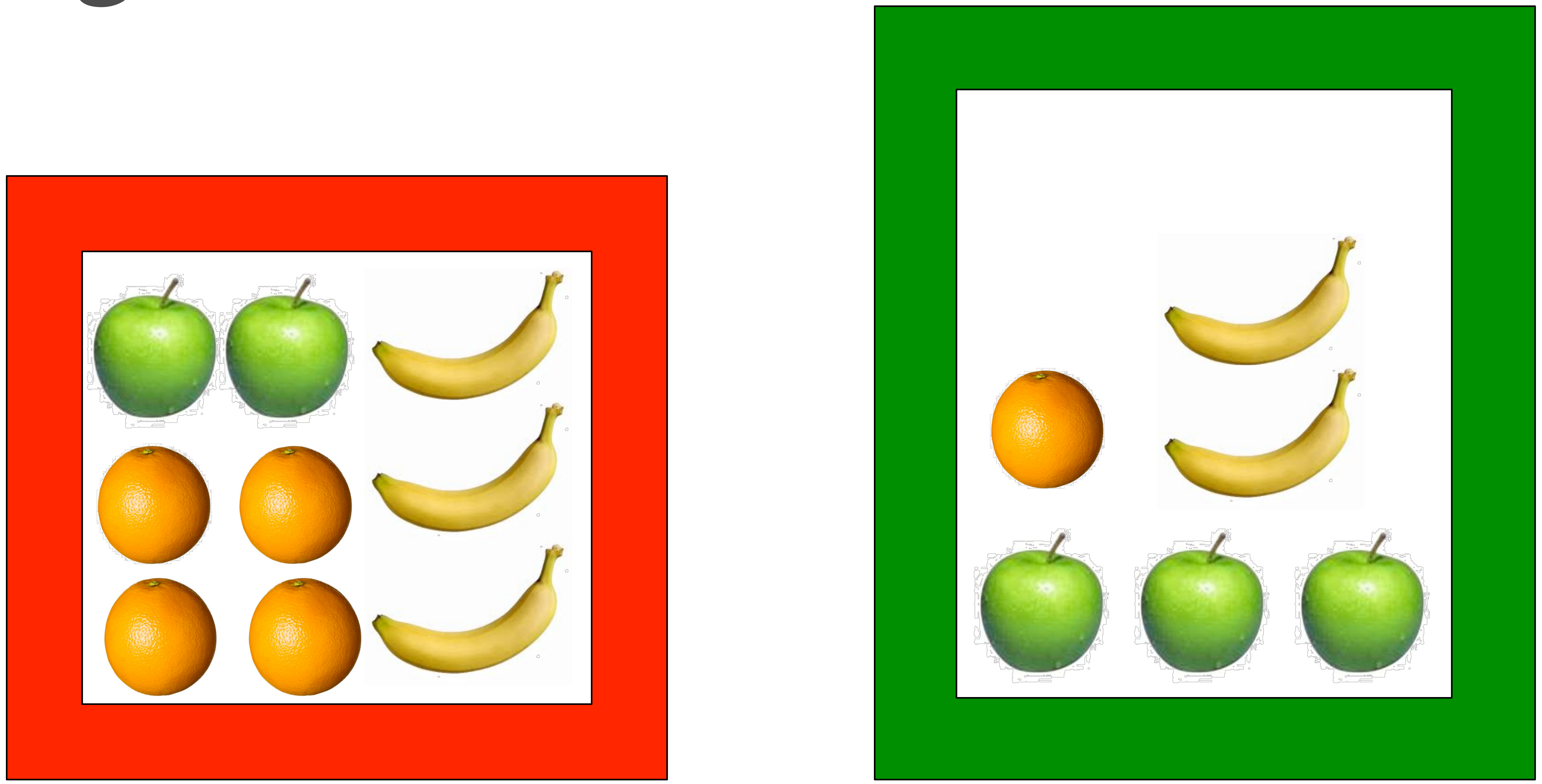
Box probabilities

- Obviously:
  - $P(B == g) = 6/10$
  - $P(B == r) = 4/10$
  - $P(\cdot) \in [0,1]$



Asking questions

- What is the probability of picking an apple?
- If we pick an orange, what is the probability that it came out of the green box?



Keeping track

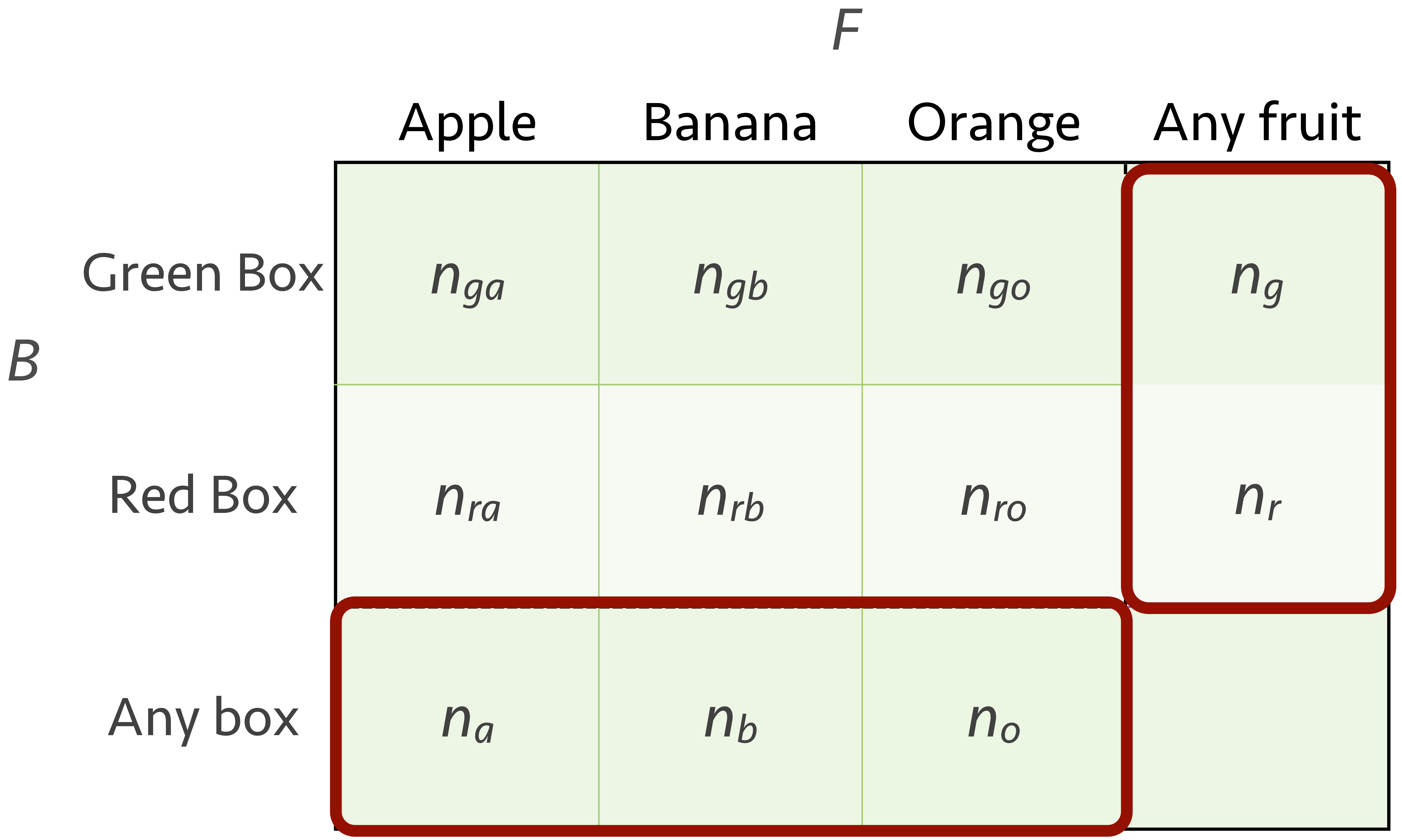
- Keep track of N experiments in a table
  - N is large, even infinite

		F			
		Apple	Banana	Orange	Any fruit
B	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	

Single variable probabilities

$$P(B == i) = n_i \,/\, N$$

$$P(F == j) = n_j \,/\, N$$



		F			
		Apple	Banana	Orange	Any fruit
B	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	

Joint probabilities

$$P(B == i, F == j) = \frac{n_{ij}}{N}$$

$$P(B == i, F == j) = P(F == j, B == i)$$

			F		
		Apple	Banana	Orange	Any fruit
B	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
		n_a	n_b	n_o	

The sum rule

$$n_i / N = \left(n_{ia} + n_{ib} + n_{io}\right) / N$$

$$P(B == i) = \sum_{\forall j} P(B == i, F == j)$$

F

B

	Apple	Banana	Orange	Any fruit
Green Box	n_{ga}	n_{gb}	n_{go}	n_g
Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	n_a	n_b	n_o	

Conditional probability

$$P(F == j \,|\, B == i) = \frac{n_{ij}}{n_i}$$

F

		Apple	Banana	Orange	Any fruit
B	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
		n_a	n_b	n_o	

The product rule

$$P(B == i, F == j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_i}\frac{n_i}{N} = P(F == j \mid B == i)P(B == i)$$

F

		Apple	Banana	Orange	Any fruit
B	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
		n_a	n_b	n_o	

The two basic rules

- Sum Rule:

$$P(X) = \sum_{Y} P(X,Y)$$

- Product Rule:

$$P(X,Y) = P(Y \mid X)P(X)$$

Bayes theorem

- From product rule & joint symmetry

$$\overset{\text{Posterior}}{P(Y \mid X)} = \frac{\overset{\text{Likelihood}}{P(X \mid Y)}\overset{\text{Prior}}{P(Y)}}{\underset{\text{Normalizing constant}}{P(X)}}$$

- Will answer most of your questions!

Independence

- If:

$$P(B == i, F == j) = P(B == i)P(F == j)$$

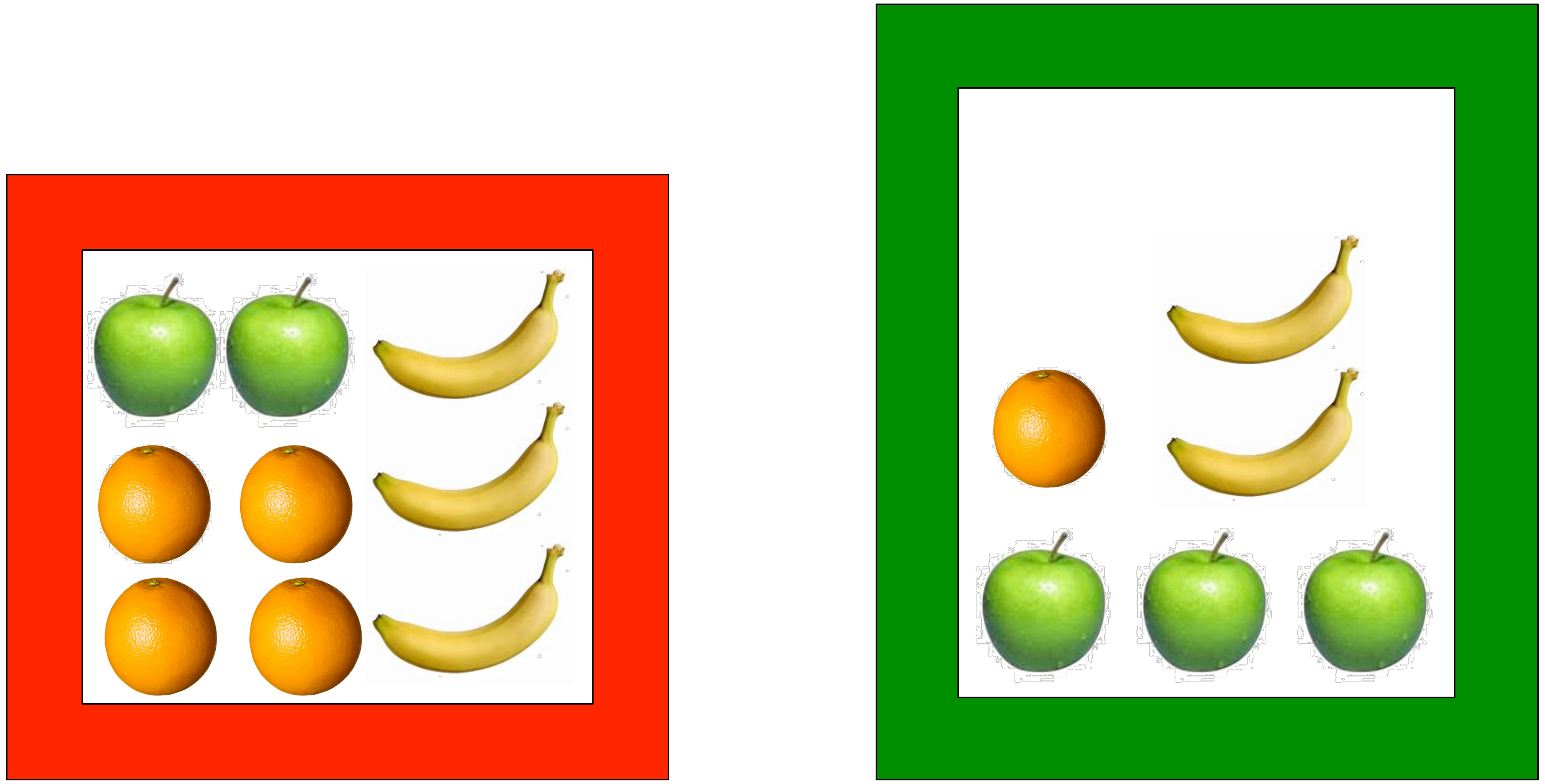
- Then B and F are independent
- Also means, via the product rule, that:

$$P(F \mid B) = P(F)$$

- If both boxes had the same fraction of fruits, then we would have independence

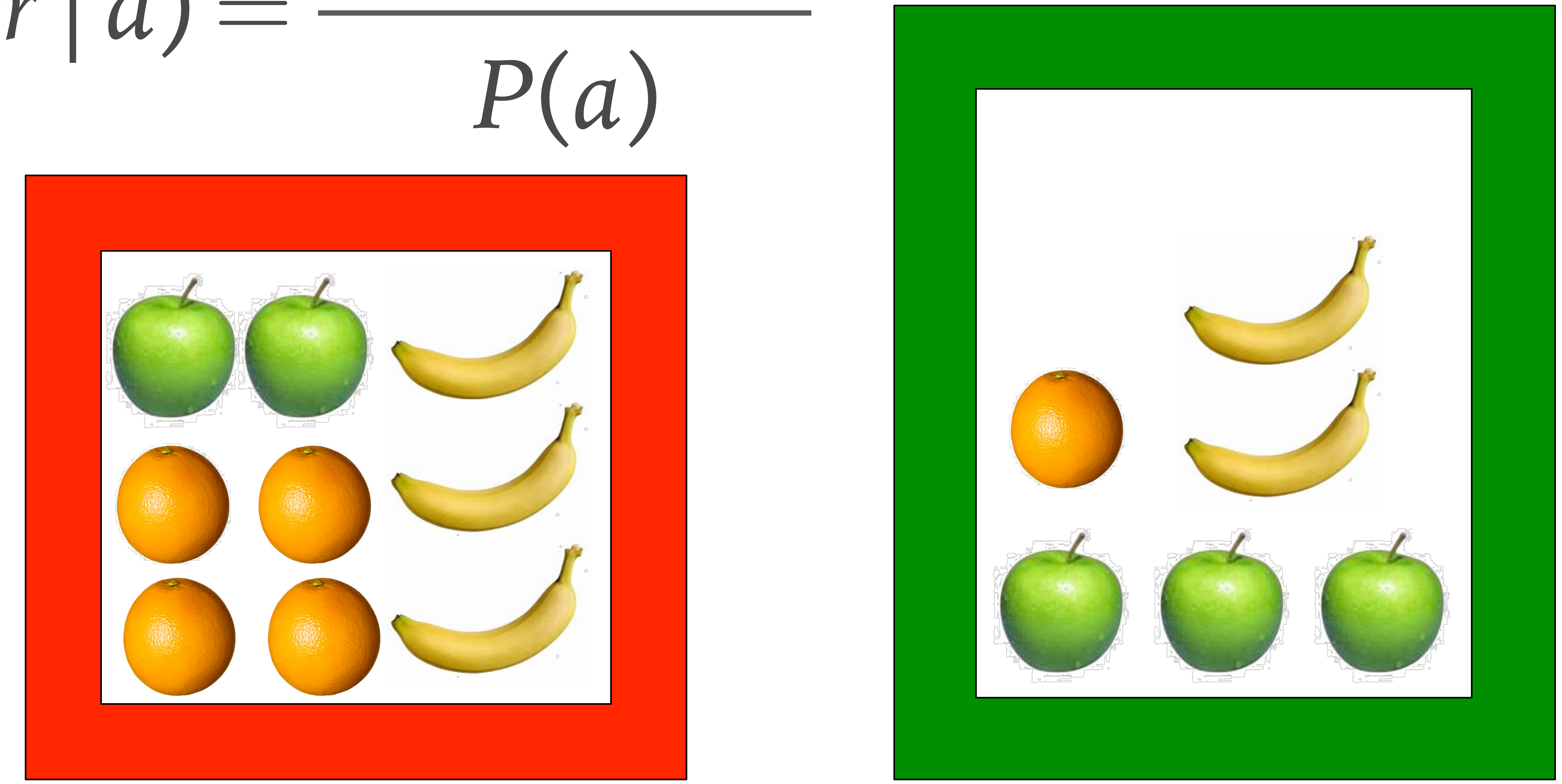
Back to the fruit

- What's the probability of picking a banana?
  - Sum rule: $P(b) = P(b,r) + P(b,g)$



Back to the fruit

- What's the probability of the red box given that I picked an apple?
  - Bayes rule: $P(r \mid a) = \dfrac{P(a \mid r)P(r)}{P(a)}$



Schools of thought

- Frequentists
  - Probabilities are interpretations of frequencies of occurrence in experiments
    - There can only be one solution!
- Bayesians
  - Probabilities are a degree of belief, not a result of a counting experiment
    - What’s the distribution of the parameter? The priors?

Why belief?

- “Will a meteor hit earth?”
  - Frequentist: Let us wait until *N* is large ...
- Using a Bayesian treatment we can find a likelihood given the evidence, not the data
  - But that requires models, priors, assumptions, ... More later

A practical application

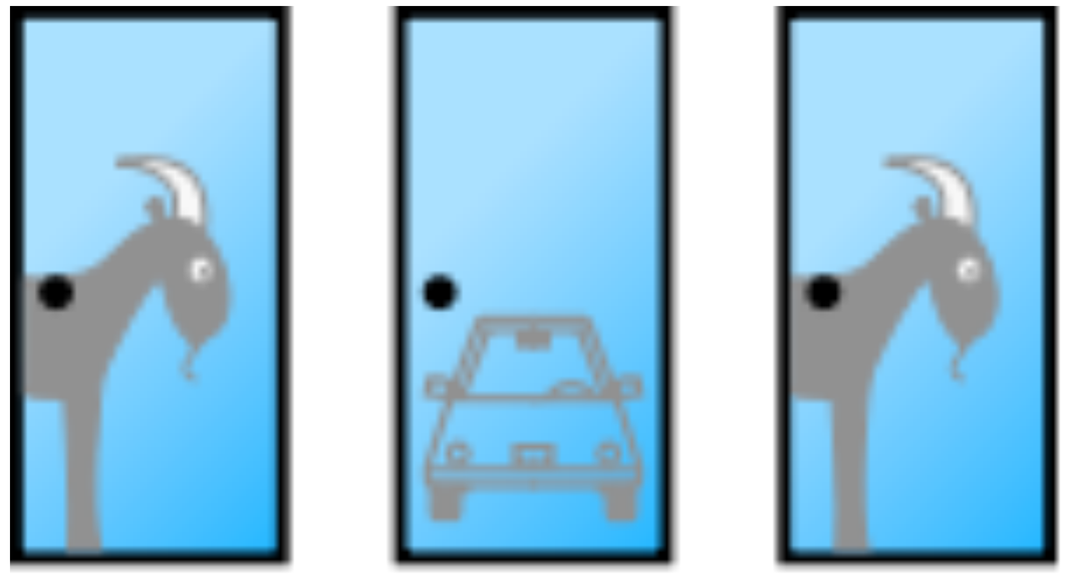
Statistics in the Real World:
The Search for the USS Scorpion

http://www.youtube.com/watch?v=U9-G-noZrwc

Getting lost? Don't worry

- Probability is super tricky
  - Even seasoned professionals get it wrong!
    - E.g. the Monty Hall problem

http://marilynvossavant.com/game-show-problem/

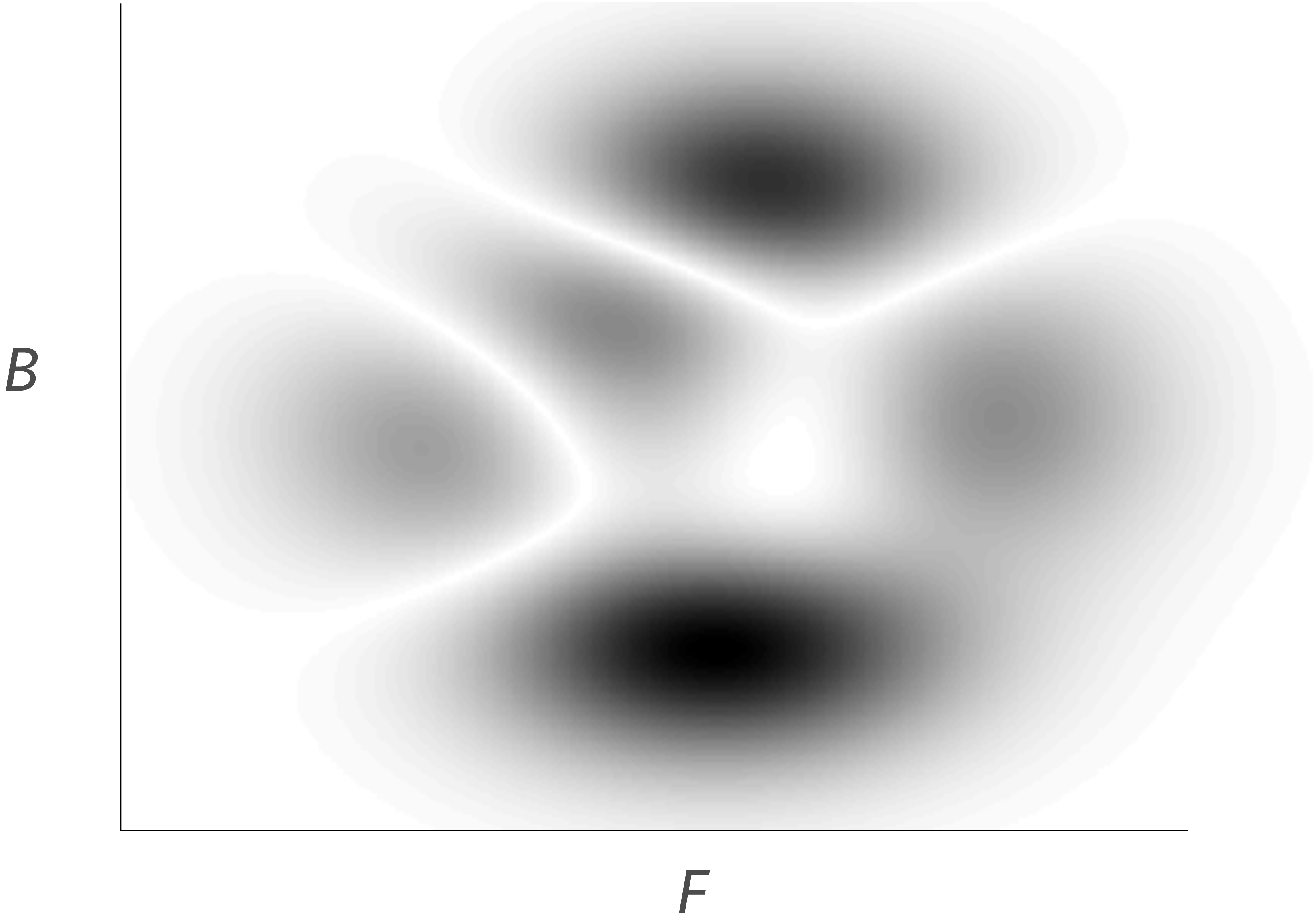


Quick answer

		Door 1	Door 2	Door 3	Outcome
Pick door 1 and switch	1st case	Car	Goat	Goat	Switch & lose
	2nd case	Goat	Car	Goat	Switch & win
	3rd case	Goat	Goat	Car	Switch & win
Pick door 1 and stay	4th case	Car	Goat	Goat	Switch & win
	5th case	Goat	Car	Goat	Switch & lose
	6th case	Goat	Goat	Car	Switch & lose

Continuous distributions

- What if we have infinite colors of boxes, and infinite types of fruit?



Same(ish) rules (harder proofs)

- Sum rule: $P(x) = \int P(x,y)\,dy$

- Product rule: $P(x,y) = P(y \mid x)P(x)$

- Bayes rule: $P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$

Some properties

- Integration to unity

$$\int_{-\infty}^{\infty} P(x) = 1$$

    - You'll be amazed how many get this wrong!

- Probabilities are real and non-negative

$$P(x) \in \mathbb{R} \qquad P(x) \geq 0$$

  - Well, they don't have to be. More on that later …

Useful operations

- Expectation: $\mathrm{E}\left(f(x)\right) = \int P(x)f(x)dx$
- Conditional expectation: $\mathrm{E}_x\left(f(x) \mid y\right) = \int P(x \mid y)f(x)dx$
- Variance: $\mathrm{var}\left(f(x)\right) = \mathrm{E}\left[\left(f(x) - \mathrm{E}\left[f(x)\right]\right)^2\right] = \mathrm{E}\left[f(x)^2\right] - \mathrm{E}\left[f(x)\right]^2$
- Covariance: $\mathrm{cov}[x, y] = \mathrm{E}_{x,y}(xy) - \mathrm{E}(x)\mathrm{E}(y)$

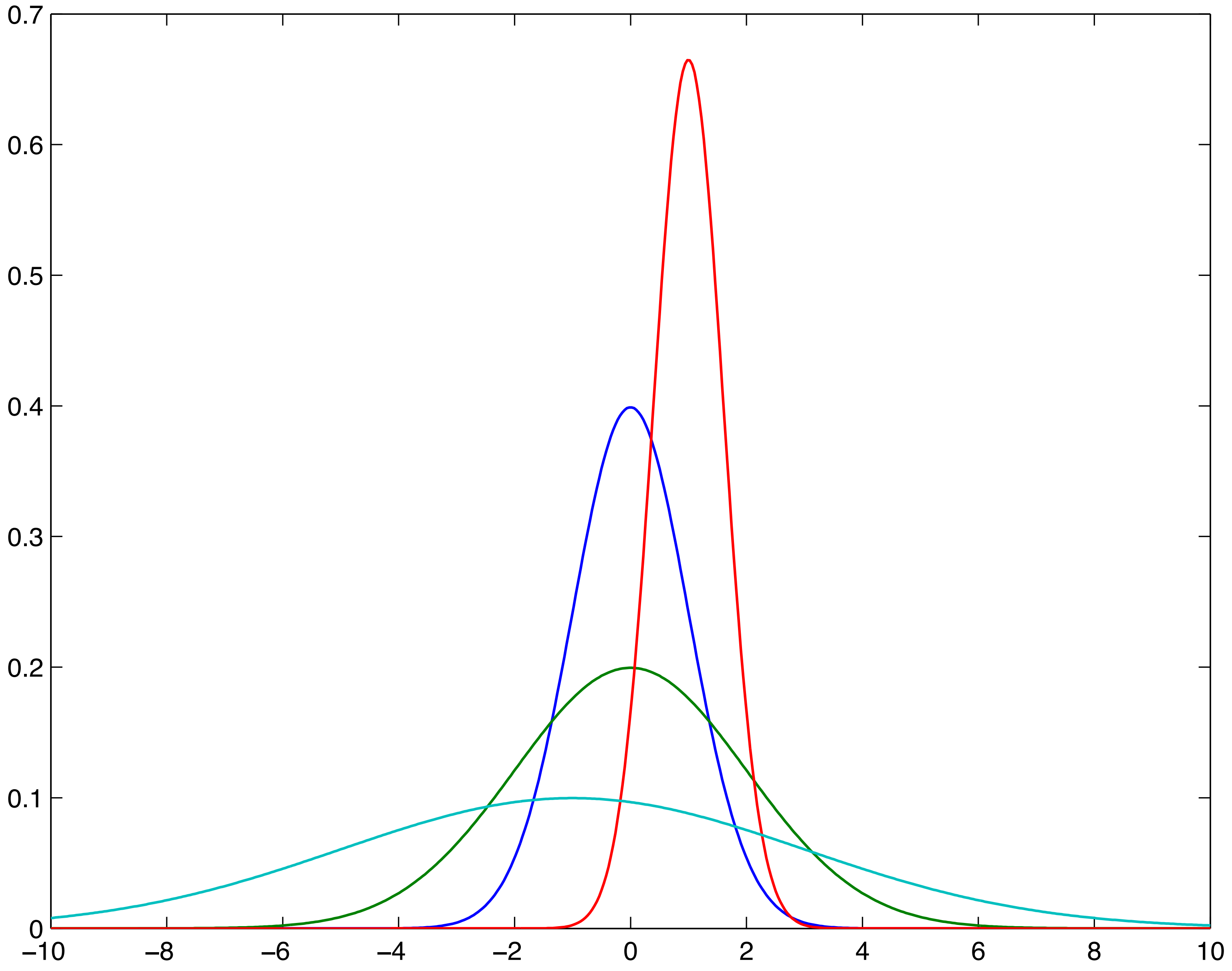
Popular distributions

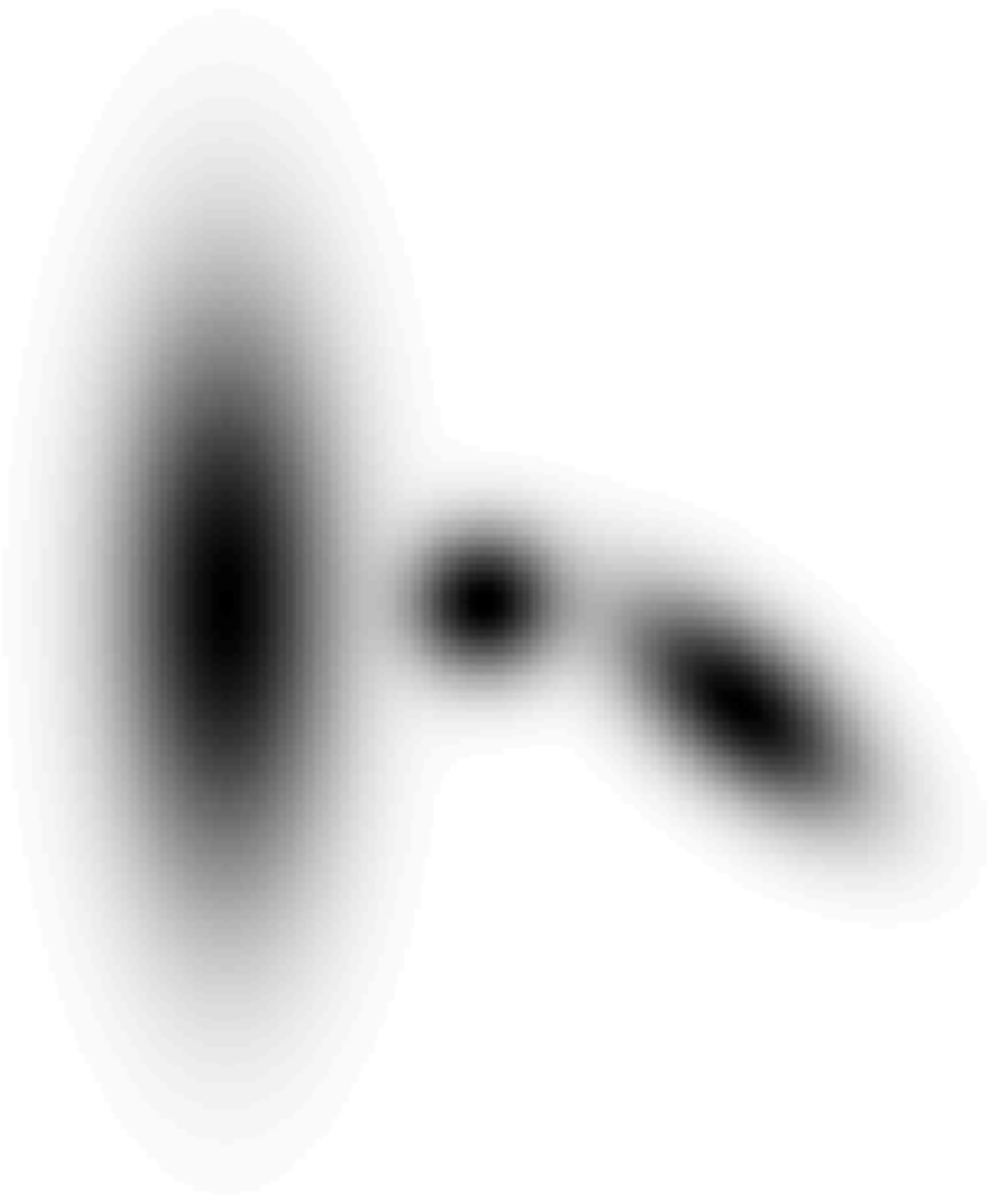
- We’ll be seeing a lot of:
  - The Gaussian
    - Used pretty much everywhere
  - The Laplacian
    - Used for sparse models
  - The Dirichlet
    - Used for compositional models
  - The Exponential Family
    - Very useful properties!

The Gaussian

- Also known as the Normal distribution or the bell curve

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi^{D}\,|\boldsymbol{\Sigma}|}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \qquad \mathbf{x}\in\mathbb{R}^{D}$$


One-dimensional Gaussians


Two-dimensional Gaussians

Why the Gaussian?

- Makes the Euclidean distance a distribution

$$\mathcal{N}(x;\mu,\sigma)=\frac{1}{(2\pi\sigma^2)^{1/2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- If you assume squared Euclidean errors, then you are using a Gaussian

The Gaussian parameters

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \qquad \mathbf{x} \in \mathbb{R}^D$$

- The mean: $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$
- The covariance: $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$
  - The mode: $\mathrm{mode}(\mathbf{x}) = \boldsymbol{\mu}$

Special case

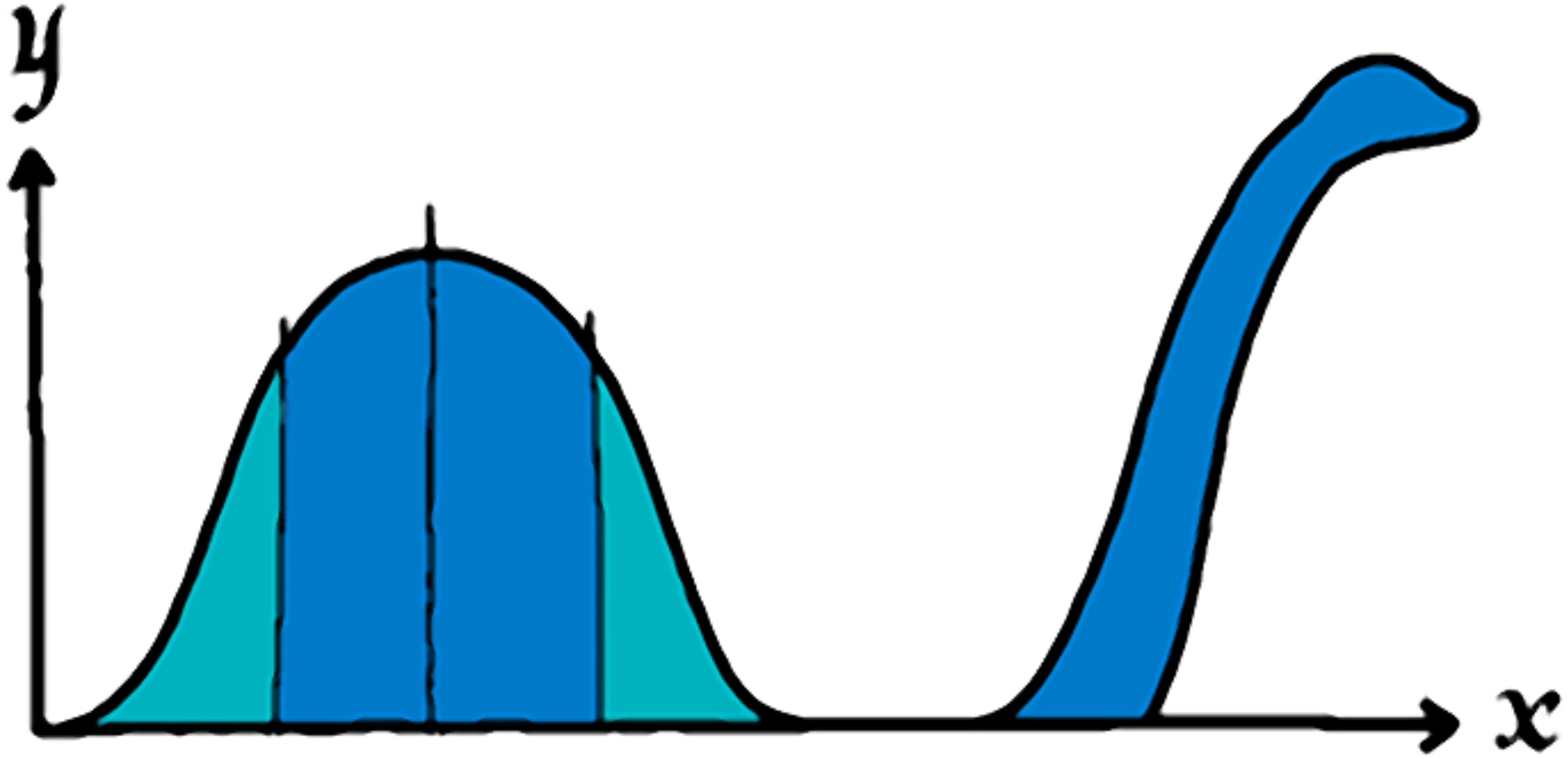


fig 1.0 The Extended Bell Curve.

– *by Tang Yau Hoong*

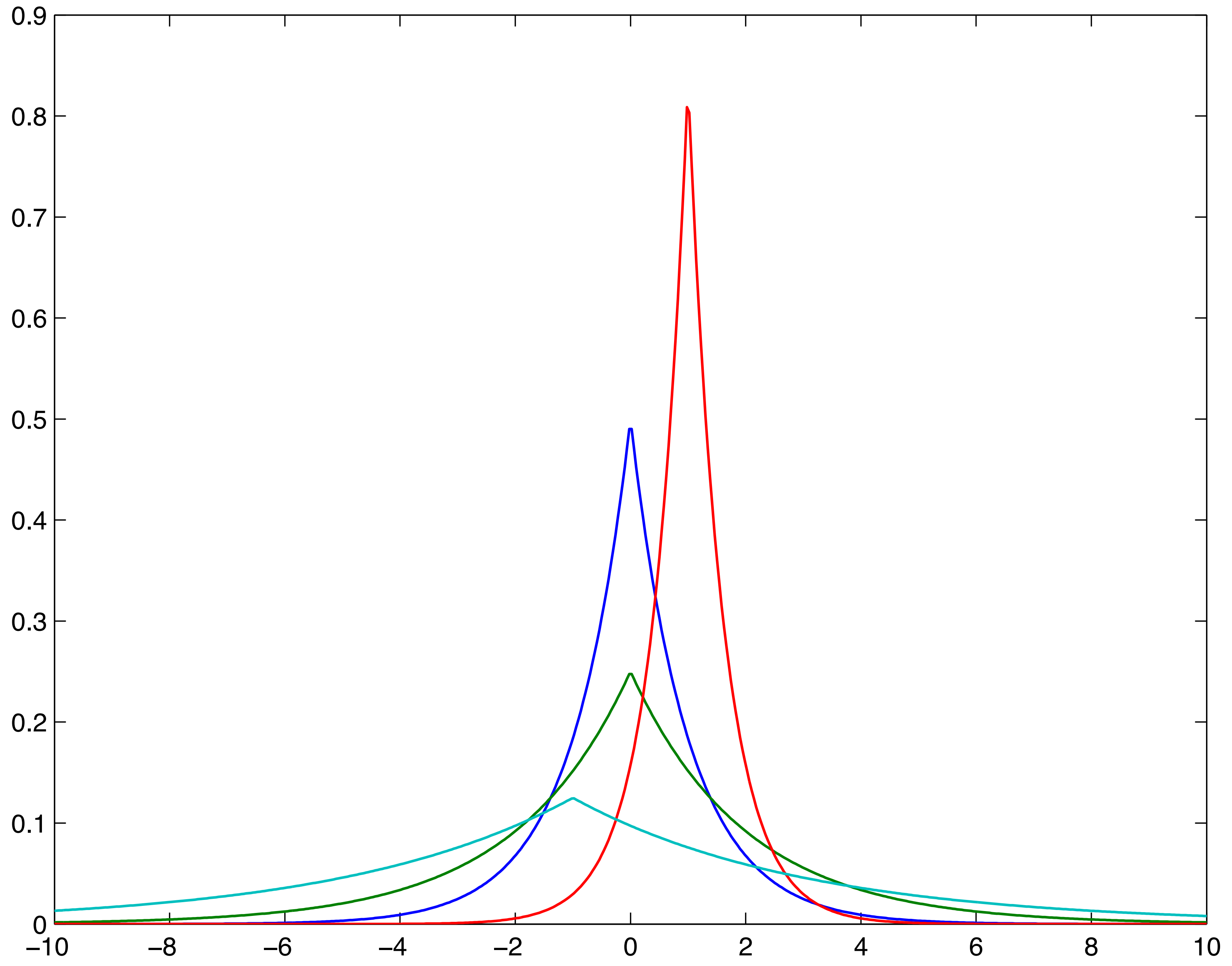
The Laplacian

- Sharper than the Gaussian
  - Uses absolute distance, not Euclidean

$$P(x;\mu,b)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$$

  - Mean: μ
  - Variance: $2b^2$
    - Mode: μ

One-dimensional Laplacians



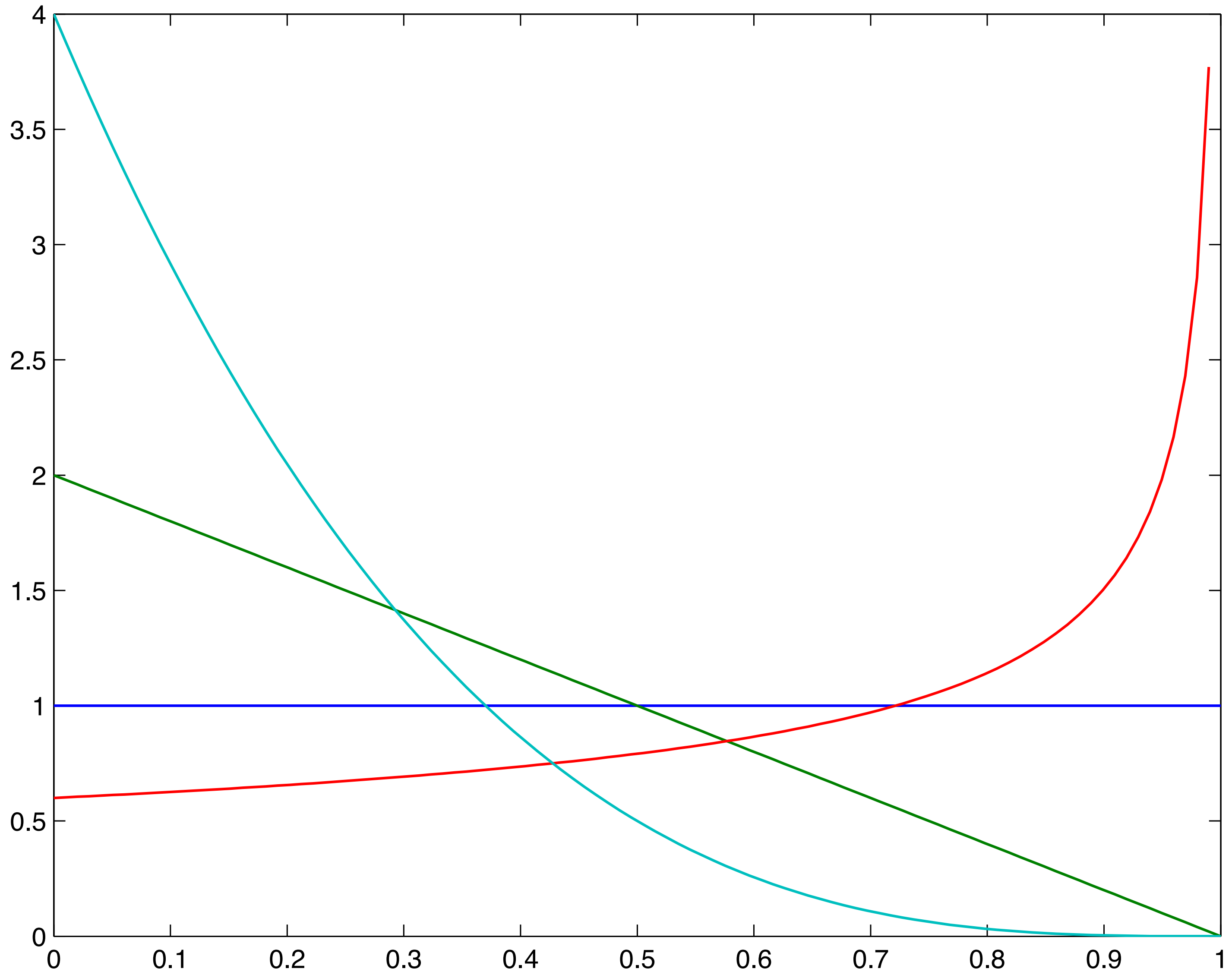
Beta/Dirichlet distributions

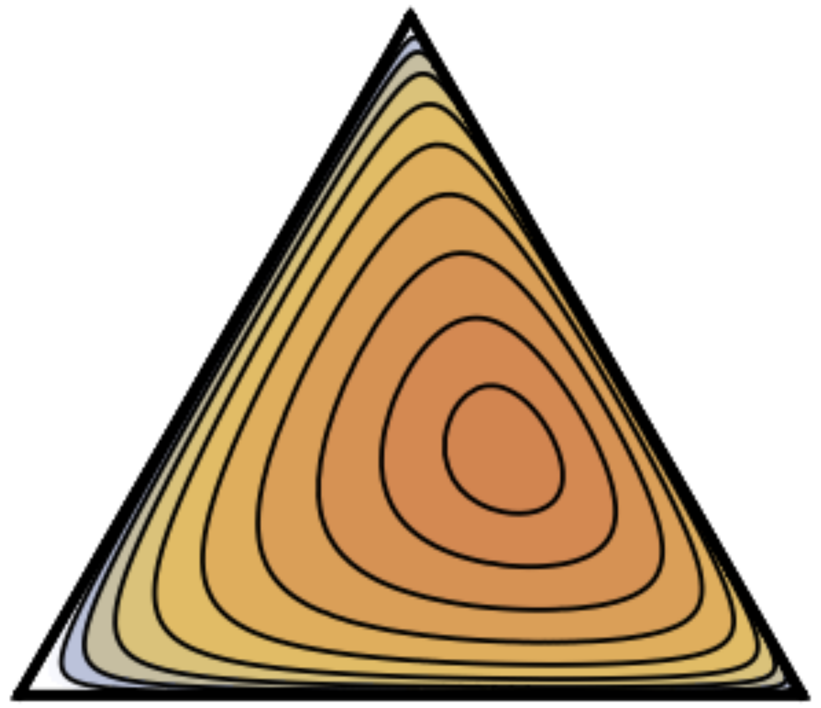
- Defined on a simplex
  - $x_1 + x_2 + x_3 + \ldots = 1$

$$P(\mathbf{x};\mathbf{a}) = \frac{\prod \Gamma(a_i)}{\Gamma\left(\sum a_i\right)} \prod x_i^{a_i - 1}$$

  - For 1D the Dirichlet is the Beta
- Mean: $\mathrm{E}[x_i] = a_i / a_0$
- Variance: $\mathrm{cov}[x_i, x_j] = \dfrac{-a_i a_j}{a_0^{\,2}(a_0 + 1)}$
- Mode: $x_i = (a_i - 1) / (a_0 - K)$

Beta distributions on the {2-1}-simplex





Dirichlet distribution on a {3-1}-simplex

The exponential family

- Any distribution that can be written as:

$$P(\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})}$$

- $\boldsymbol{\eta}$ contains the natural parameters
- $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$
- $g(\boldsymbol{\eta})$ is just for normalization

Gaussian example

$$P(\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})}$$

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \qquad h(\mathbf{x}) = (2\pi)^{-1/2}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \mu / \sigma^2 \\ -1 / 2\sigma^2 \end{bmatrix}, \quad g(\boldsymbol{\eta}) = (-2\boldsymbol{\eta}_2)^{1/2} e^{\boldsymbol{\eta}_1{}^2/4\boldsymbol{\eta}_2}$$

$$P(\mathbf{x};\mathrm{h}) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}\mu^2} = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Why this mess???

- Allow us to see a broader picture

- Exponential distributions have convenient properties
  - Sufficiency
    - You won't need more parameters for more data
  - Conjugate priors
    - Make life easy when we perform parameter estimation (more later)

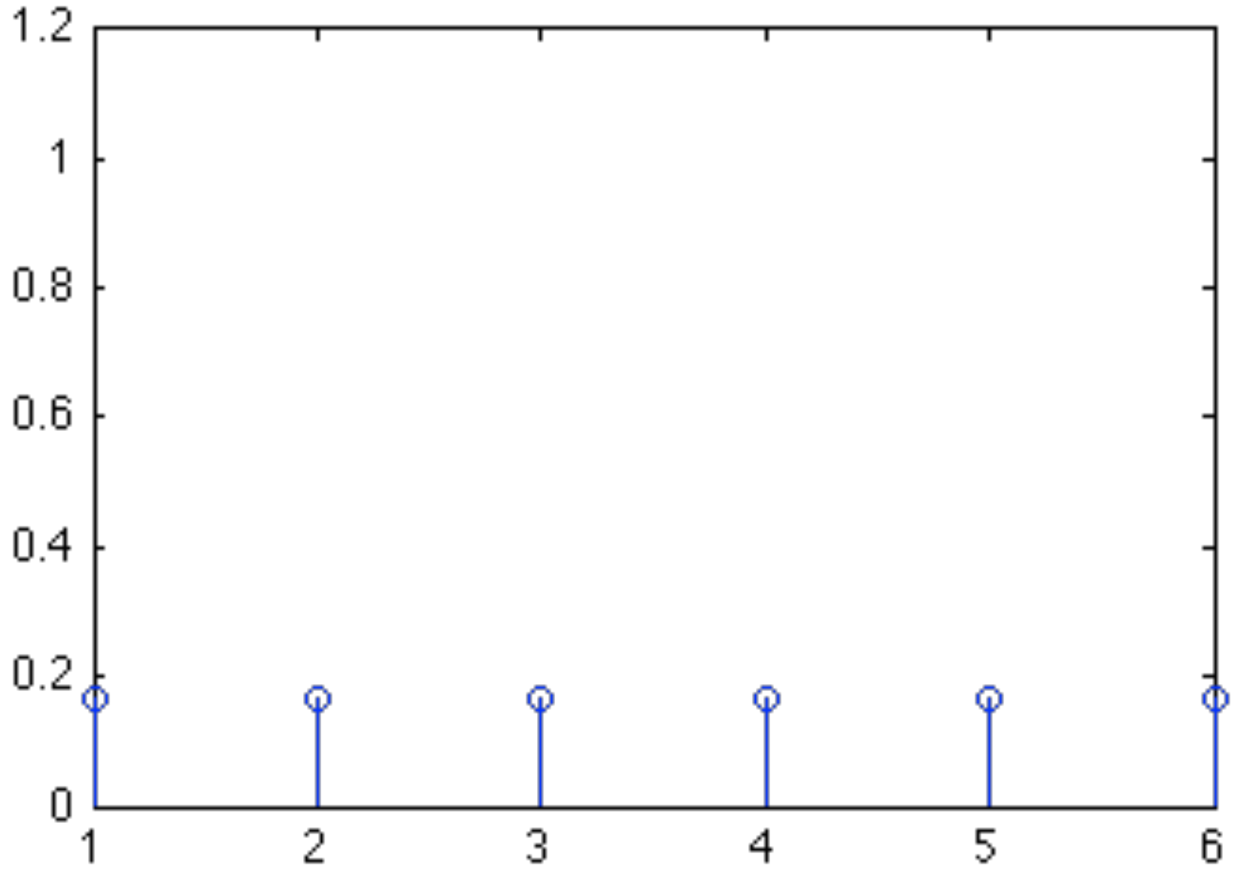
Information theory

- Entropy

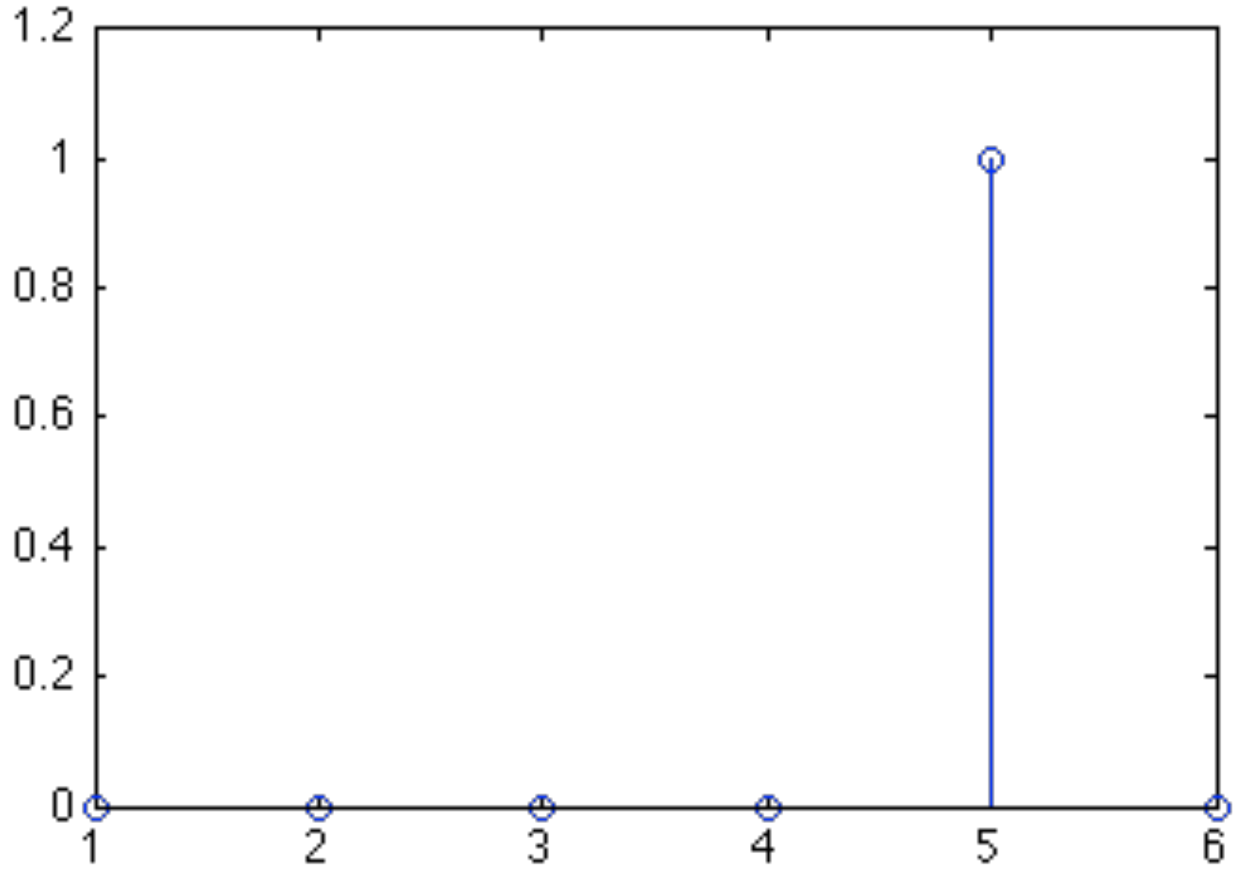
$$H(x) = -\int P(x)\log P(x)\,dx \quad \text{or} \quad -\sum_x P(x)\log P(x)$$

$$H(x,y) = -\int\int P(x,y)\log P(x,y)\,dx\,dy \quad \text{or} \quad -\sum_x\sum_y P(x,y)\log P(x,y)$$

  - A measure of information in a distribution



A fair die, H = 1.79
There is a lot of uncertainty
therefore more information



A heavily biased die, H = 0
no message to convey

Information theory

- Mutual information
  - Measures amount of shared information

$$I(x,y) = H(x) + H(y) - H(x,y)$$

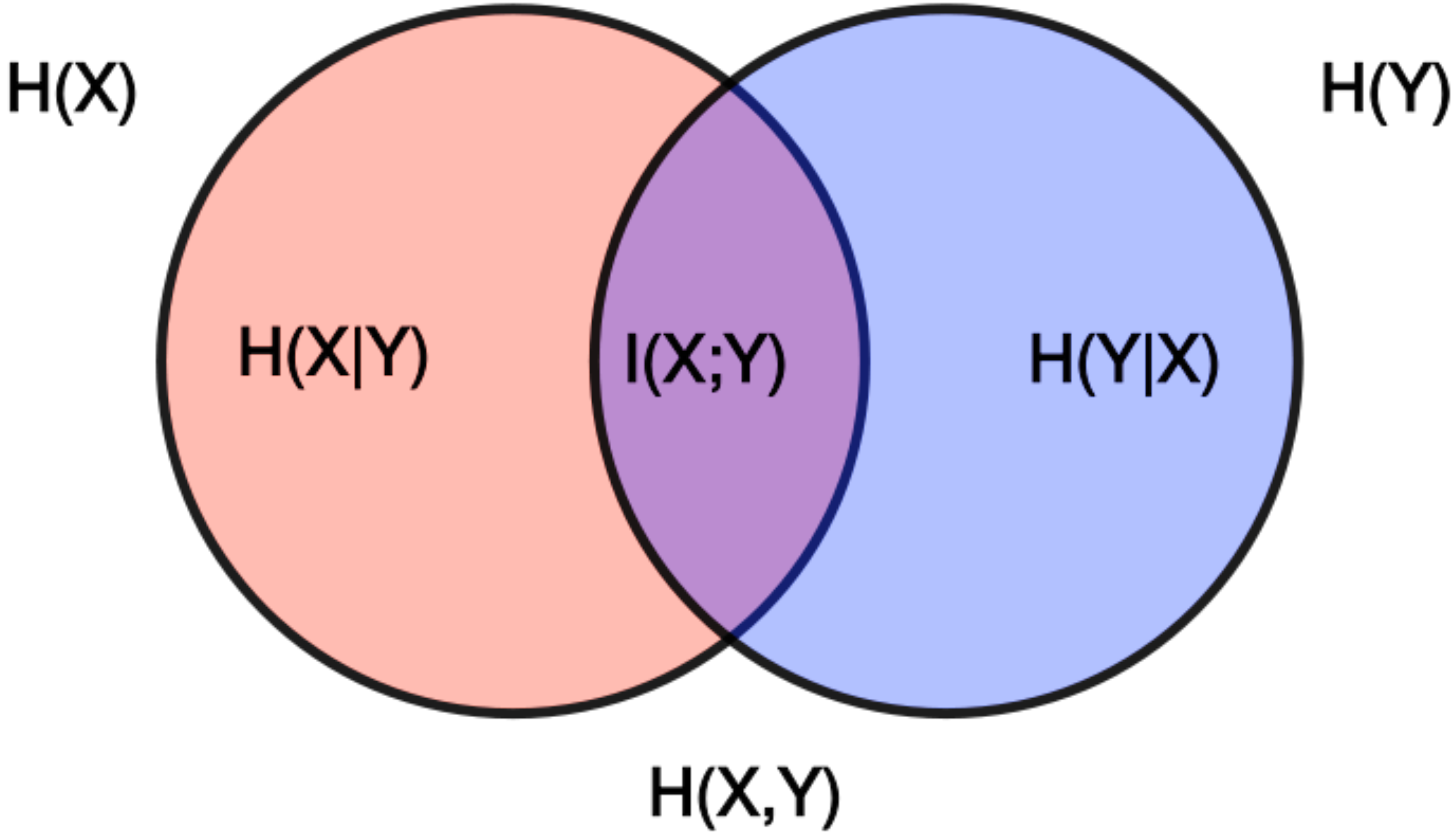
  - If 0 then x,y are independent
- Kullback-Leibler divergence
  - a pseudo-distance for distributions

$$D(p\,||\,q) = \sum p_i \log \frac{p_i}{q_i} \quad \text{or} \quad \int p(x) \log \frac{p(x)}{q(x)} dx$$

$$D(P(x,y)\,||\,P(x)P(y)) = I(x,y)$$

  - If 0 then p and q are the same

Entropy types



Parameter estimation

- So what do we do with distributions?
  - We like to explain data with them
- To do so we need parameter estimation
  - Find the distribution parameters that result in explaining the observed data best
  - Various ways to go about it

Parameter estimation

- Given some independent samples:

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

- and a model:

$$P(\mathbf{X}; \theta)$$

- Find the parameters θ

Maximum likelihood

- The overall likelihood is:

$$P(\mathbf{X};\theta) = P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N;\theta) = \prod_i P(\mathbf{x}_i;\theta)$$

- We want to find:

$$\theta_{ML} = \arg\max_{\theta} \prod_i P(\mathbf{x}_i;\theta)$$

- We can use straightforward solving

Maximum likelihood

- Set the derivative to zero:

$$\frac{\partial \prod_i P(\mathbf{x}_i;\theta)}{\partial \theta} = 0$$

- Go to the log domain to remove product:

$$\frac{\partial \log \prod_i P(\mathbf{x}_i;\theta)}{\partial \theta} = \sum_i \frac{\partial \log P(\mathbf{x}_i;\theta)}{\partial \theta} = \sum_i \frac{1}{P(\mathbf{x}_i;\theta)} \frac{\partial P(\mathbf{x}_i;\theta)}{\partial \theta} = 0$$

- Substitute your P and solve

Example

- Mean of Gaussian distributed data
  - Define the model:

$$P(\mathbf{x};\mu,\sigma^2)=\prod_{i=1}^{N}\mathcal{N}(\mathbf{x};\mu,\sigma^2)$$

  - Form log-likelihood:

$$\log P(\mathbf{x};\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_n-\mu)^2-\frac{N}{2}\log\sigma^2-\frac{N}{2}\log 2\pi$$

  - Set derivative to zero and solve:

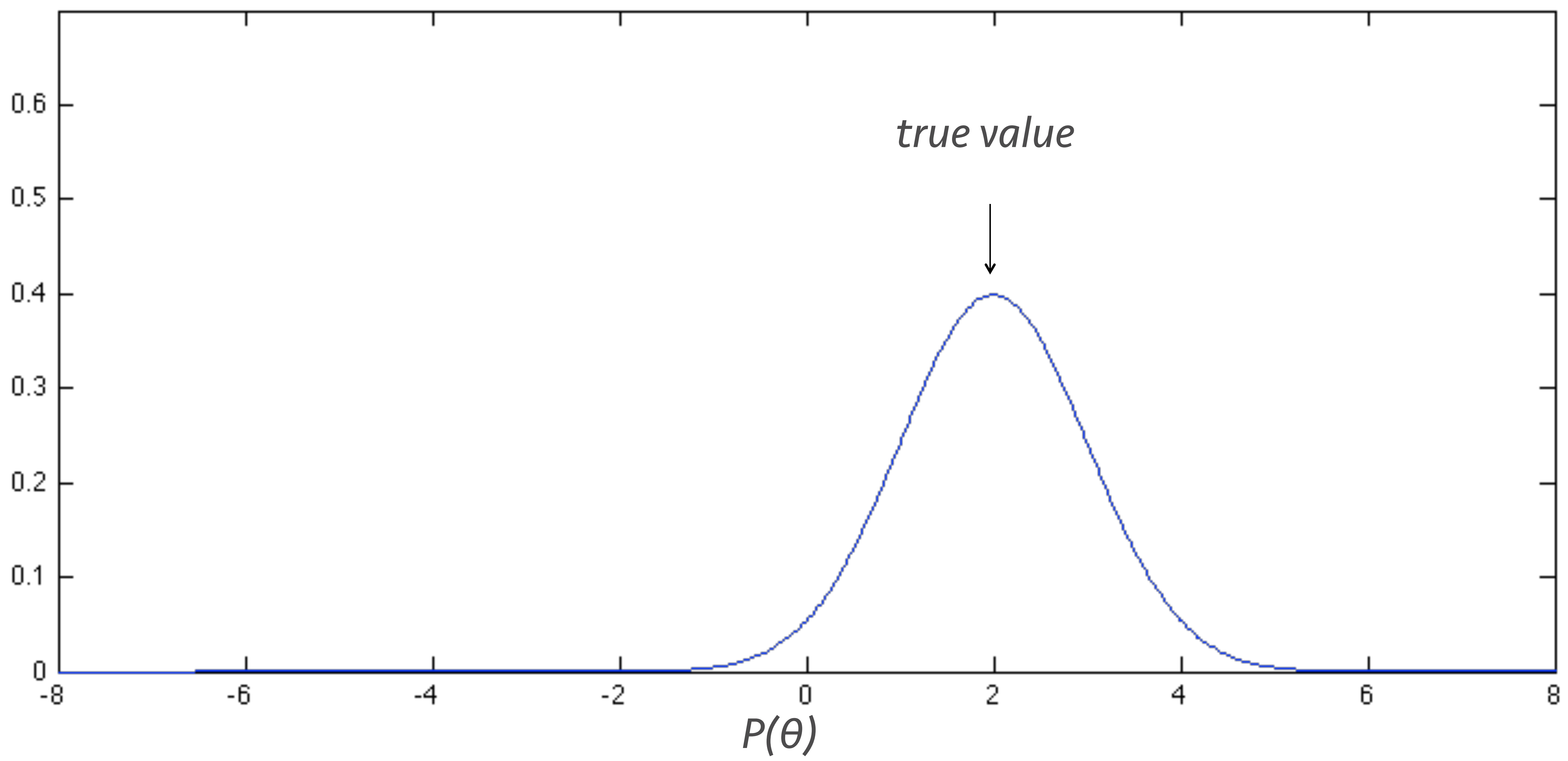
$$\frac{\partial\log P(\mathbf{x};\mu,\sigma^2)}{\partial\mu}=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\frac{\partial(x_n-\mu)^2}{\partial\mu}=0\Rightarrow\mu=\frac{1}{N}\sum_{i=1}^{N}x_i$$

Wait a minute!

- All that to prove the obvious?
- Yes, it is tedious
  - In many cases the answer will be obvious
    - But keep in mind that looks might be deceiving!
- In other cases the answer will not be easy
  - Requiring numerical/approximate optimization

A couple of ML properties

- The ML estimate is (usually) asymptotically Gaussian distributed and: $\lim_{N\to\infty} \mathrm{E}[\theta_{\mathrm{ML}}] = \theta_{\mathrm{true}}$ $\lim_{N\to\infty} \mathrm{E}\left[\left\|\theta_{\mathrm{ML}} - \theta_{\mathrm{true}}\right\|^2\right] = 0$



Maximum a posteriori (MAP)

- Sometimes we have a prior belief
  - E.g. we believe the answer should be close to a value
  - Maximum likelihood doesn't incorporate that
  - MAP does

- Same setup as before but in addition to $P(x;\theta)$ we also have a $P(\theta)$

MAP estimation

- We use Bayes' theorem and we now maximize:

$$P(\theta \mid \mathbf{x}) = \frac{P(\theta)P(\mathbf{x} \mid \theta)}{P(\mathbf{x})}$$

- The denominator is constant so we only have to maximize the numerator:

$$\theta_{MAP} = \arg\max_{\theta} P(\theta)P(\mathbf{x} \mid \theta)$$

- Same story as before …

MAP estimation example

- Estimate the mean, but use a prior:

$$P(x;\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x;\mu,\sigma^2), \quad P(\mu;\mu_0,\sigma_\mu^2) = \mathcal{N}(\mu,\mu_0,\sigma_\mu^2)$$

- Take log, differentiate, solve:

$$\frac{\partial}{\partial\mu} \log \prod_{i=1}^{N} P(x_i \mid \mu) P(\mu) = 0$$

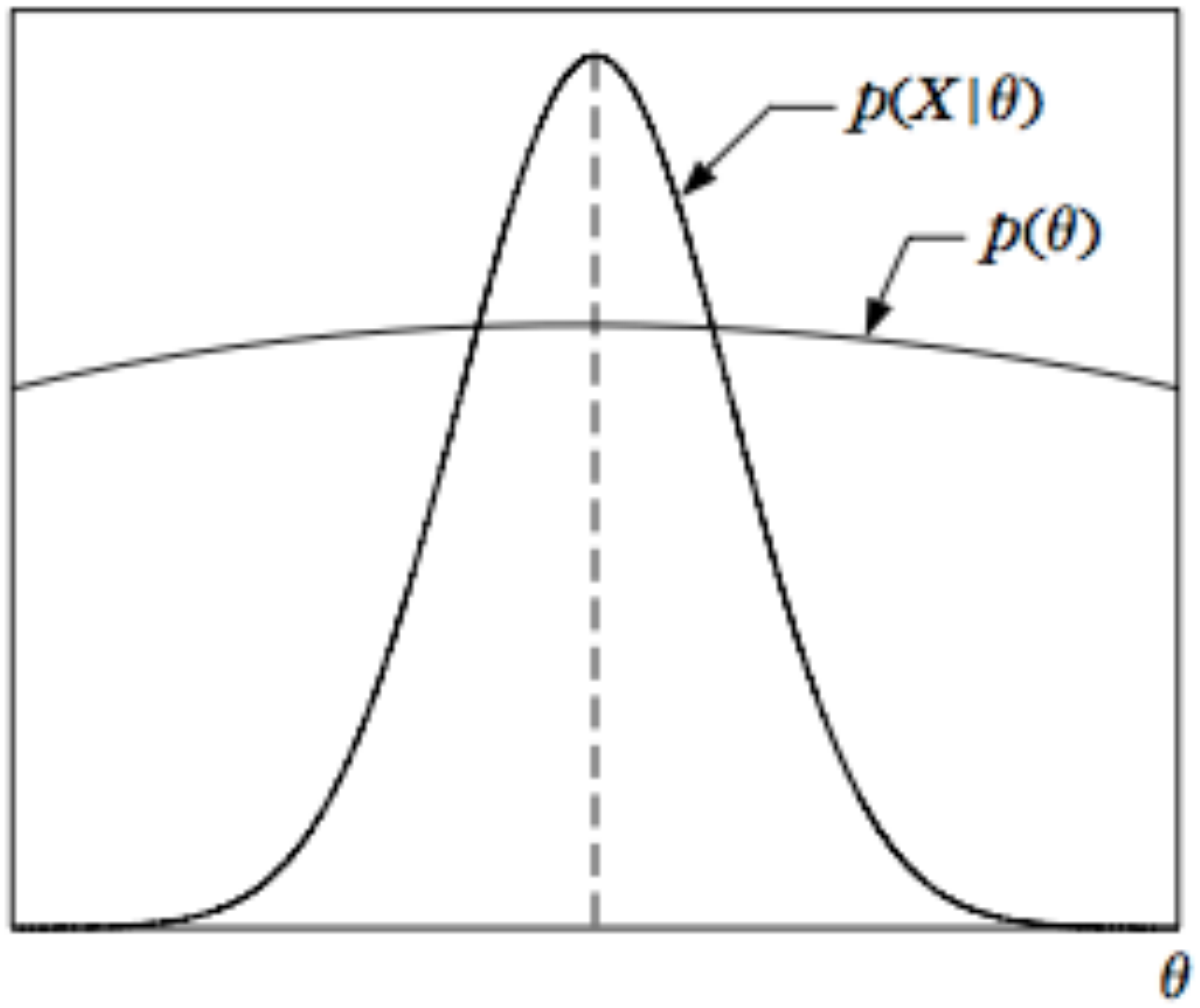
$$\sum_{i=1}^{N} \frac{1}{\sigma^2}(x_i - \mu) - \frac{1}{\sigma_\mu^2}(\mu - \mu_0) = 0$$

$$\Rightarrow \mu_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{N} x_i}{1 + \frac{\sigma_\mu^2}{\sigma^2} N}$$

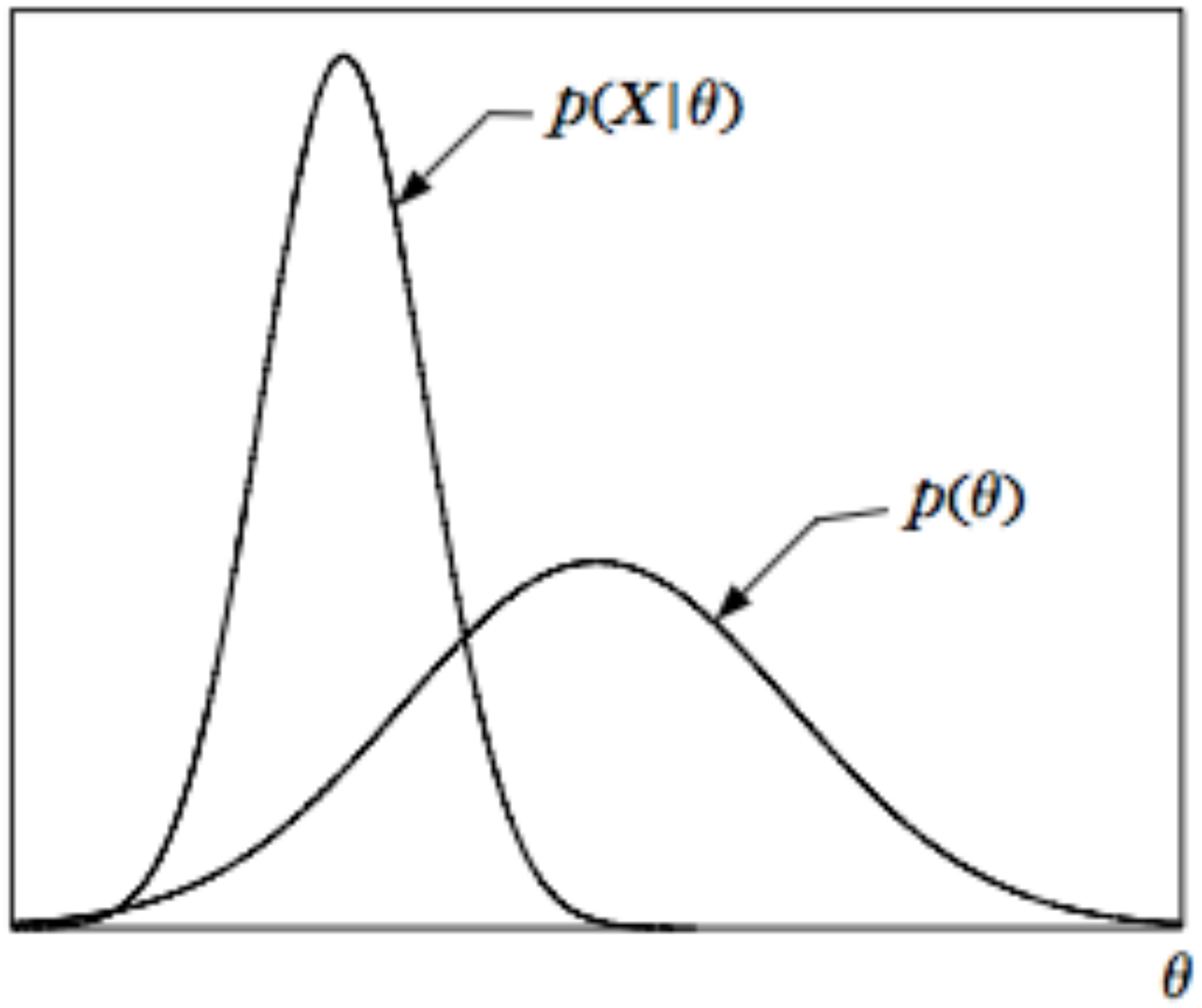
MAP vs. ML

- If $P(\theta)$ is uniform then MAP == ML
  - Otherwise they will most likely not coincide

ML and MAP will be the same

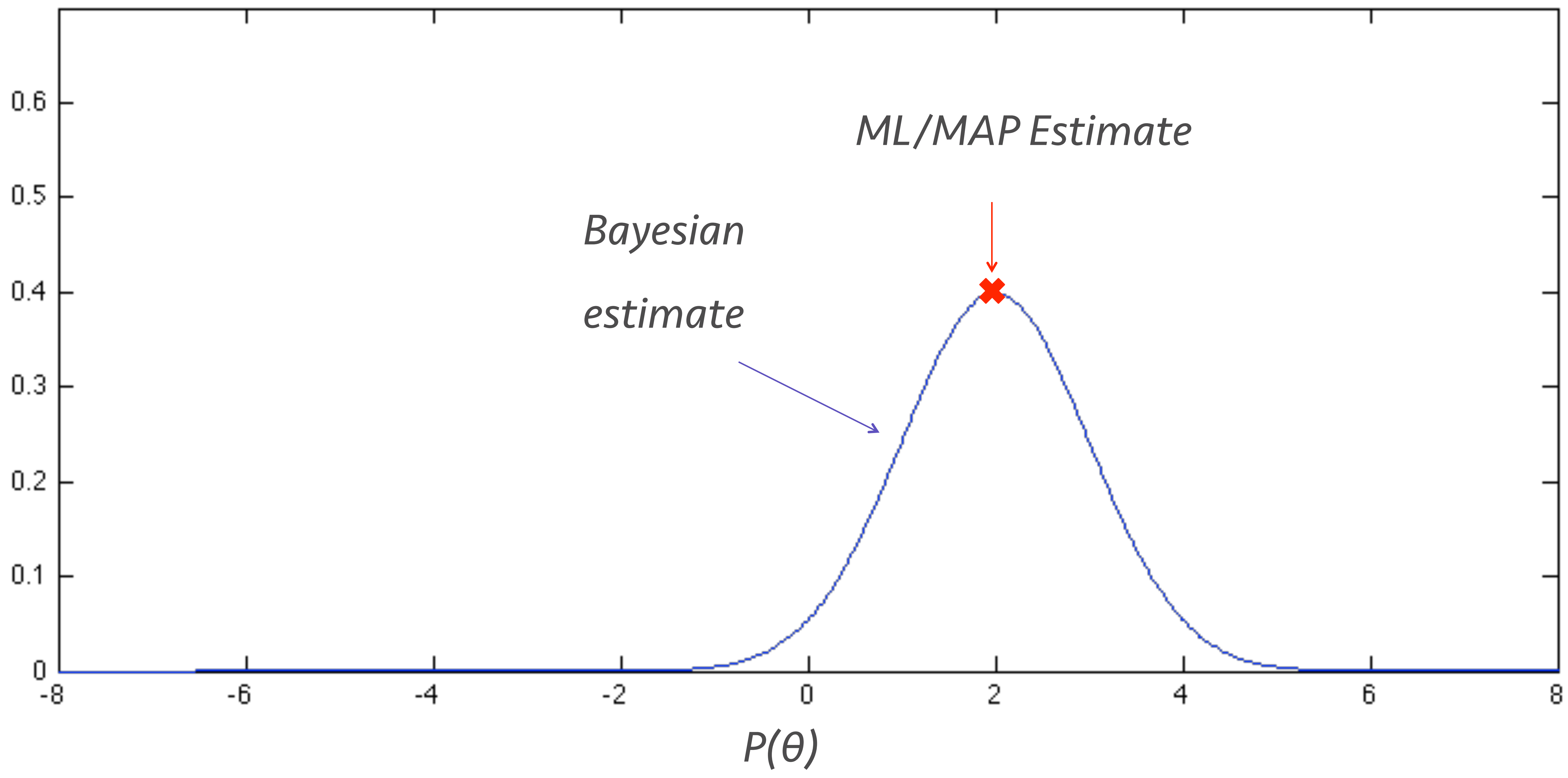


ML and MAP will be different



Bayesian inference

- Bayesian inference doesn't care about the optimal value, it cares about it's distribution



Example estimation

- Same setup as in the MAP case:

$$P(\mathbf{x};\mu,\sigma^2)=\prod_{i=1}^{N}\mathcal{N}(\mathbf{x};\mu,\sigma^2),\qquad P(\mu;\mu_0,\sigma_\mu^2)=\mathcal{N}(\mu,\mu_0,\sigma_\mu^2)$$

- We now find the distribution of the mean:

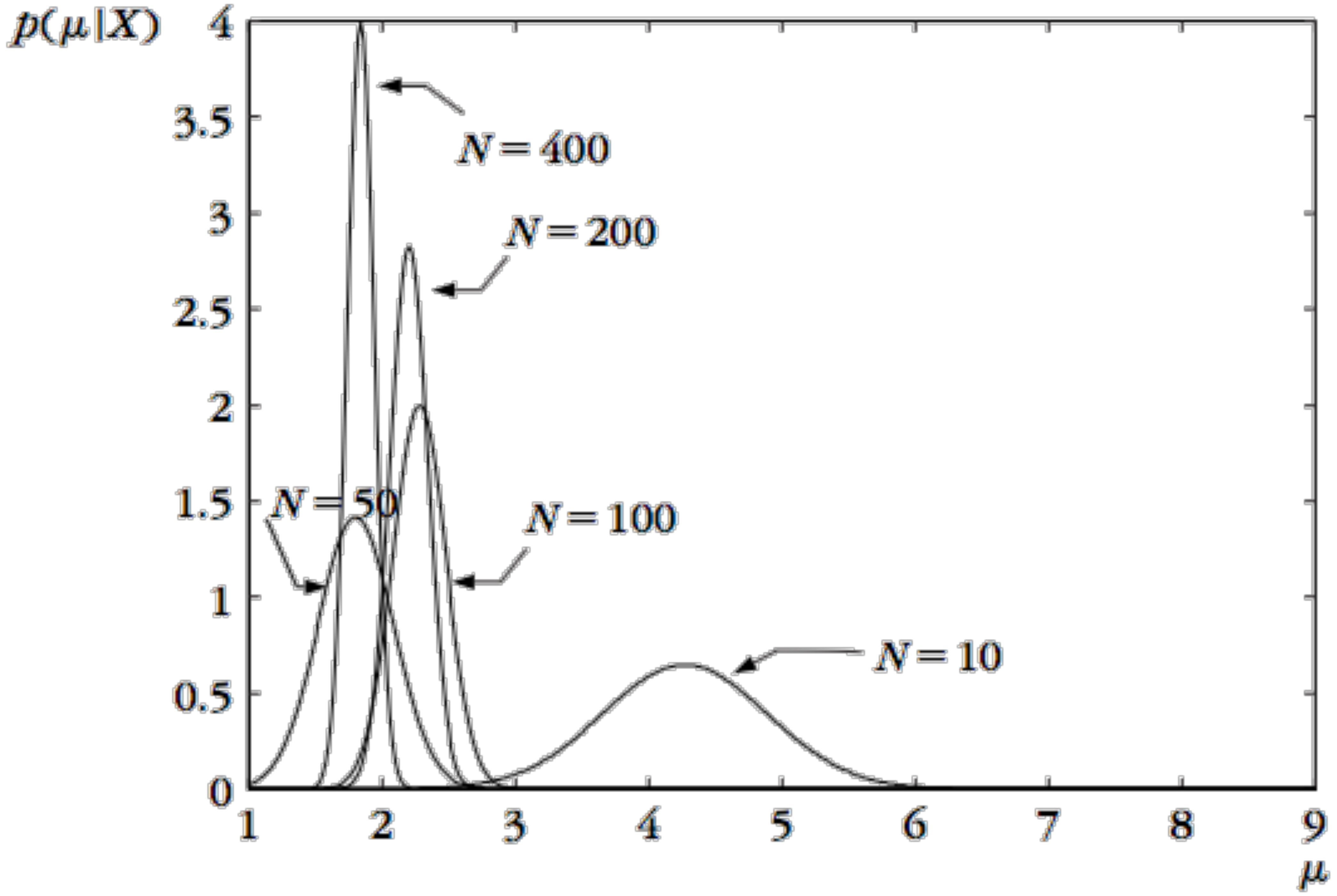
$$P(\mu\,|\,\mathbf{X})=\frac{P(\mathbf{X}\,|\,\mu)P(\mu)}{P(\mathbf{X})}=\ldots=\mathcal{N}(\mu,\mu_N,\sigma_N^2)$$

$$\mu_N=\frac{N\sigma_0^2\mathrm{E}[\mathbf{x}]+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2},\qquad \sigma_N^2=\frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}$$

  - Which is also Gaussian!

Obtaining the estimate

- For different values of N we obtain a different distribution of the parameter we estimate
  - The bigger the N the more sharp the distribution



And that was a clean case

- Often the distributions don’t work out

- We resort to numerical solutions
  - Usually sampling (Monte Carlo, etc.)

Other methods

- Maximum entropy estimation
  - Choose model that maximizes entropy
    - Least committal approach
- Expectation-Maximization
  - Useful for mixture models
  - We'll cover in detail later

Recap

- Probability
  - sum/product/Bayes rules
- Distributions
  - Gaussian, Laplacian, Dirichlet
- Information theory
  - Entropy, Mutual Info, KL divergence
- Parameter estimation
  - ML, MAP, Bayesian

Too much information?

- You are not supposed to master all this
  - We will be encountering these ideas later
  - This lecture should serve as a reference

Some more reading

- Get textbook from class page
  - UIUC network access only
- Probability basics
  - Appendix 1 of textbook
- Parameter estimation
  - Section 2.5 of textbook

Next week

- Signals refresher
  - “All of DSP in a lecture”